Comparing Complex Diagnoses: A Formative Evaluation of the Heart Disease Program

Hamish S F Fraser[1, 2], William J Long[1], Shapur Naimi[2]
[1]MIT Lab for Computer Science, Cambridge, MA
[2]Tufts-New England Medical Center, Boston, MA

Background. The Heart Disease Program (HDP) is a large diagnosis program developed to diagnose cardiac disease, particularly hemodynamic dysfunction and heart failure. It utilizes a pseudo Bayesian belief network augmented with mechanisms to reason about severity and timing of diseases. Cases are entered through HTML forms; the output is a set of hypotheses listing all likely diseases, with explanations of how each item is justified by the input. The HDP has already undergone two evaluations in a laboratory setting with retrospectively collected cases. Here we describe initial results and analysis methods of a prospective observational study in realistic clinical settings.

Methods. Cases are selected and entered by physicians independently of the researchers. Once a case is entered, the physicians are asked for their own diagnosis, which are kept separate from program's results. For comparison "Gold Standard" (GS) diagnoses are obtained from detailed chart review with particular emphasis on investigations such as echocardiography and cardiac catheterisation. In addition case summaries without the diagnoses are given to independent cardiologists who generate their own differential diagnosis. Comparison is then made between the HDP diagnosis, and the standard diagnoses obtained from the entering physician, the cardiologists and the GS. Two measures are used to compare the HDP's diagnoses and the three standards. (1) *Sensitivity*; defined as the proportion of items in a standard diagnosis that match the HDP output. (2) *Positive predictive value* (PPV); defined as the proportion of items in the HDP diagnosis present in the standard. *Specificity* is difficult to measure due to the large number of possible diagnoses most of which do not occur in the average case (values would be around 95%).

Results. The table shows values for Sensitivity and PPV on the first 27 cases. The program's performances against the three standards and cross validation of the standards are shown. It must be noted that in assessing the performance of the HDP and doctors against the GS the data on which the program is carrying out the analysis is much less than is available at discharge time. The values for Sensitivity and PPV therefore appear relatively low, but are comparable to those obtained in a recent study of medical diagnosis programs[4]. In that study Comprehensiveness a measure equivalent to Sensitivity used here, ranged from 0.25 – 0.38 (HDP 0.5). Relevance a measure similar to PPV, ranged from 0.19 – 0.37 (HDP 0.31). HDP results are the means of the first 3 columns in the table.

Conclusions. The program was upgraded after the first 60 cases in response to this evaluation. 130 cases have now been entered; full follow-up should be available by fall 1997.

Comparison of the HDP and three Different Standard Diagnosis Lists for 27 Cases

Comparison	HDP -> Cardiologist	HDP -> Gold Standard	HDP -> Physician	Cardiologist-> Gold Standard	Physician-> Gold Standard
Sensitivity	0.43	0.46	0.62	0.34	0.19
PPV	0.33	0.37	0.24	0.41	0.61

References

1) Long W. Medical Diagnosis Using a Probabilistic Causal Network. Applied Artificial Intelligence. 1989; 3:367-83.

2) Long WJ, Naimi S, Criscitiello MG. Evaluation of a New Method for Cardiovascular Reasoning. JAMIA 1994; 1:127-141.

3) Long W J, Fraser H S F, Naimi S, A Web Interface for the Heart Disease Program, Proceedings of the 1996 AMIA Annual Fall Symposium, October 26 to 30, 1996. Editor James J. Cimino MD.

4) Berner E S, Webster, G D Shugerman, A A Jackson, J R et al. Performance of Four Computer-Based Diagnostic Systems, NEJM, 330, No. 25, 1792-1796

1091—8280/97/$5.00 © 1997 AMIA, Inc.